Research Article | Vol. 13, No. 1 / 1 Jan 2022 / Biomedical Optics Express 284

Biomedical Optics EXPRESS

Deep-3D microscope: 3D volumetric microscopy of thick scattering samples using a wide-field microscope and machine learning

Bowen Li,[1] Shiyu Tan,[2] Jiuyang Dong,[3] Xiaocong Lian,[1] Yongbing Zhang,[4] Xiangyang Ji,[1,5] and Ashok Veeraraghavan[2,6]

[1]*Department of Automation & BNRist, Tsinghua University, Beijing, China*
[2]*Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005, USA*
[3]*Tsinghua Shenzhen International Graduate School, Shenzhen, China*
[4]*Harbin Institute of Technology (Shenzhen), Shenzhen, China*
[5]*xyji@tsinghua.edu.cn*
[6]*vashok@rice.edu*

Abstract: Confocal microscopy is a standard approach for obtaining volumetric images of a sample with high axial and lateral resolution, especially when dealing with scattering samples. Unfortunately, a confocal microscope is quite expensive compared to traditional microscopes. In addition, the point scanning in confocal microscopy leads to slow imaging speed and photobleaching due to the high dose of laser energy. In this paper, we demonstrate how the advances in machine learning can be exploited to "teach" a traditional wide-field microscope, one that's available in every lab, into producing 3D volumetric images like a confocal microscope. The key idea is to obtain multiple images with different focus settings using a wide-field microscope and use a 3D generative adversarial network (GAN) based neural network to learn the mapping between the blurry low-contrast image stacks obtained using a wide-field microscope and the sharp, high-contrast image stacks obtained using a confocal microscope. After training the network with widefield-confocal stack pairs, the network can reliably and accurately reconstruct 3D volumetric images that rival confocal images in terms of its lateral resolution, z-sectioning and image contrast. Our experimental results demonstrate generalization ability to handle unseen data, stability in the reconstruction results, high spatial resolution even when imaging thick (~40 microns) highly-scattering samples. We believe that such learning-based microscopes have the potential to bring confocal imaging quality to every lab that has a wide-field microscope.

1. Introduction

High-throughput, high-resolution, high-contrast microscopy techniques, that do not damage tissue are critical for multiple domains including scientific imaging, pathology, medical imaging, and in-vivo imaging. The current workhorse of microscopy is a wide-field microscope and every science lab, pathologist's office and hospital/clinic in every corner of the globe likely have access to one. While wide-field microscopes have truly been democratized, they are for the most part only suited to image the surface of thin samples. 3D volumetric imaging, especially with scattering tissue samples is a rapidly growing need that wide-field microscopes cannot address.

Existing techniques for 3D volumetric imaging in scattering samples such as confocal microscopy [1,2], two-photon microscopy [3–8], and light-sheet microscopy [9–14] all rely on more complex optics and illumination designs that end up being prohibitively expensive for many parts of the world. The question we ask in this paper is "Can the revolutionary advances in machine learning over the last decade be exploited to make the data acquired from conventional

#444488 https://doi.org/10.1364/BOE.444488
Journal © 2022 Received 22 Oct 2021; revised 2 Dec 2021; accepted 3 Dec 2021; published 10 Dec 2021

Biomedical Optics EXPRESS

wide-field microscopes rival 3D volumetric data acquired using confocal microscopes — even when imaging thick scattering samples?"

AI-enhanced fluorescence microscopy. Machine learning and artificial intelligence have made revolutionary advances in the last decade and have completely transformed a variety of applications all the way from autonomous vehicles to medical diagnostics. These revolutionary advances have been the result of (a) new deep neural network architectures that are highly over-parameterized, (b) datasets acquired to "teach" these networks, and (c) efficient algorithms both for training and for testing. This AI revolution has also begun to have significant impacts on microscopy, biological and scientific imaging. The key advantage is that, with no or little modification to conventional microscopes, AI techniques can provide additional capabilities such as high resolution [15–19], high SNR [20,21], fast acquisition [22–24], auto-focusing [25,26], and cross-modality [27–30].

Over the last few years researchers have developed early examples of deep convolutional neural networks to enhance axial resolution and imaging contrast of wide-field images [31–35]. Specifically, Zhang et al. [31] first successfully transformed wide-field images to background reduced Structured Illumination Microscopy (SIM) images using a 2D neural network. However, their demonstration was restricted to samples with relatively simple structures and showed limited enhancement. Although they subsequently demonstrated more complicated mice whole-brain SIM images reconstruction [33], the 2D convolutional network structure confined the application to physically sectioned thin tissue samples. Wu et al. proposed two more sophisticated Generative Adversarial Network (GAN) [36,37] based deep networks to predict confocal images from a single in-focus wide-field input [30] and sparse wide-field scans [35], respectively. However, they only testified this idea using 1.6 microns thickness BPAEC (Bovine Pulmonary Artery Endothelial Cells) sample or a limited z-span of C. *elegans* sample (about 8 microns).

Here, in our paper, we take a significant leap compared to these related works. While all of the above approaches were limited in thin sample thickness/simple structures, we aim to perform true three dimensional reconstruction in thick scattering samples (demonstrating at least 5× thicker samples than prior works) with more complex structures. In the thick scattering sample, the images captured by a wide-field microscope are more severely deteriorated by the scattering background than thin sample. As shown in Fig. 1, the contrast, computed by sharpness ratio between the signal region (the cell structure) and background region (scattering and noise) of wide-field images, at different axial layers in a 38-microns neuron slice can be as much as 5× lower than the contrast in a 3-microns MCF10A thin sample. Cellular structures are submerged in the noisy background in the thick sample, which brings us more difficulties on wide-field to confocal image prediction using a deep neural network. The key technical insight that allows us to move from thin samples to thick scattering samples, is the idea that in thick samples there is a lot of information cross-talk between and across sections (z-stack) and 2D neural networks are sub-optimal for capturing the structure of these complex interactions. Therefore, we develop a 3D network structure that allows us to learn statistical relationships across the entire thick sample, enabling high-resolution, high-contrast 3D reconstructions over the entire volume.

In particular, we propose a GAN-based three-dimensional convolutional neural network (WFCON-Net) that can digitally predict confocal z-stack images from measurements of a wide-field microscope. We believe this technology will allow widely accessible wide-field microscopes to capture 3D volumetric image datasets of thick scattering samples – with a quality comparable to (but slightly worse than) a confocal microscope. Our work is inspired and motivated by the recent work [30] but with three significant contributions. First, we propose a 3D convolutional network to leverage the inter-layer connections other than 2D blocks, to utilize the stronger crosstalks between different layers in thick/dense tissue samples. This strengthens the learning power to successfully recover fine structures under high magnification with stronger scattering backgrounds. Second, we add a photo-realistic VGG loss to preserve image high-frequency

Biomedical Optics EXPRESS

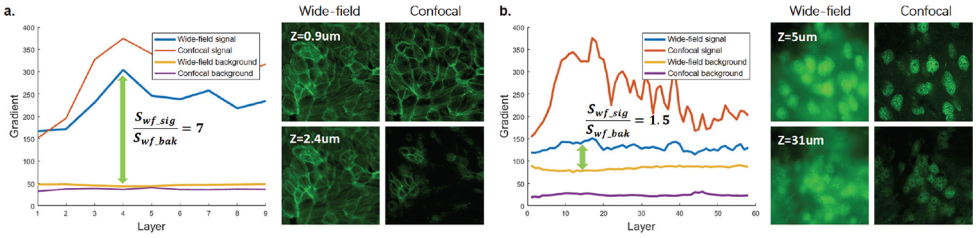

Fig. 1. Sharpness analysis for thin/thick samples. The sharpness of the signal region and the background region of the images at different axial depths are plotted. The sharpness is computed as the image gradient of the signal or background regions. **a)**. The sharpness curves for the wide-field and confocal images of a 3 microns MCF10A thin sample. **b)**. The sharpness curves for the wide-field and confocal images of a 38 microns neuron thick sample. In the thin sample, the wide-field images are less influenced by the background noise and thus have higher signal-to-noise contrast, while the signals in the thick sample are severely deteriorated by the background noises.

details. Third, we propose a 3D tailored registration technique – point spread function (PSF) based registration to accurately align cross-modality wide-field and confocal stack pairs under high noise disturbance. Furthermore, WFCON-Net can estimate dense confocal z-stacks from fewer wide-field z-scans, and thereby has the potential for further reducing the sample acquisition time. Also, we show that WFCON-Net has a good generalization ability to unseen sample data. Therefore, with the proposed method we can digitally obtain high-resolution confocal z-stacks using a typical wide-field microscope, without sacrificing the imaging depth, speed, resolution, or field of view (FOV). In summary, by using the GAN-based 3D convolutional neural network, together with the VGG loss and the 3D tailored registration technique, we succeeded in recovering true 3D confocal fluorescence of thick scattering samples from enormously degraded wide-field input.

2. Methods

Consider a wide-field microscope imaging a thick scattering tissue sample. Typically the images obtained will suffer from low-contrast and blur associated with both in-plane and out-of-plane scattering making images (especially beyond the first 10 microns of tissue) practically unusable. Even so, there are significant degrees of freedom in a wide-field microscope that one can take advantage of. The focus setting of the microscope can be slowly changed from the top to the bottom of the tissue sample to obtain a low-contrast image stack. Each image can be reasonably approximated as a linear combination of light from many layers of a 3D tissue sample. The question we ask is whether this is sufficient information to de-multiplex and recover a sharp, high-contrast volumetric image of the sample. In particular, we wish to leverage deep learning techniques and use a deep generative adversarial network to learn a mapping between the blurry, low-contrast z-stack obtained using a wide-field microscope and a sharp, high contrast 3D volume imaged using a confocal microscope.

WFCON-Net architecture. WFCON-Net is a 3D GAN-based deep neural network (Fig. 2), it consists of two parts: a generator and a discriminator. The generator is trained to generate the fake samples (the predicted confocal images) from the inputs (the wide-field images) as real samples (the true confocal images) as possible, making the discriminator believe that the fake samples are the same as real samples. The discriminator is on the contrary trained to distinguish the real samples from the fake samples produced by the generator. We denote the loss of the generator and the discriminator as L^G_{loss} and L^D_{loss}. In the training stage, the parameters of the generator and the discriminator are updated alternatively by minimizing these two losses alternatively. In the

Biomedical Optics EXPRESS

inference stage, only the generator is needed to produce whole 3D confocal-like images from 3D wide-field inputs, within a single forward pass. The use of the discriminator adds GAN losses in L^G_{loss} and L^D_{loss} for training the generator and discriminator. It encourages the generator to predict confocal z-stacks with high accuracy, providing a good match to ground truth images. More specifically, L^G_{loss} and L^D_{loss} are defined as:

$$\begin{aligned} L^G_{loss} &= [D(G(x)) - 1]^2 + \lambda L_{L1}(G(x), y) + \zeta L_{VGG}(G(x), y) \\ L^D_{loss} &= [D(G(x))]^2 + [D(y) - 1]^2 \end{aligned} \tag{1}$$

where x refers to the wide-field z-stack images, y refers to the corresponding true sharp confocal images. G and D denote the generator and the discriminator. The first term of L^G_{loss} and L^D_{loss} are GAN losses, we use least-square GAN loss for the stability of training [38]. Besides, we add L1 loss and perceptual VGG loss [39] as additional regularization to penalize the artifacts caused by the GAN framework, where λ and ζ are the corresponding weights. The use of perceptual loss encourages high-quality, high-resolution predicted confocal images. In this paper, we set $\lambda = 2$, $\zeta = 0.01$ for all the experiments.

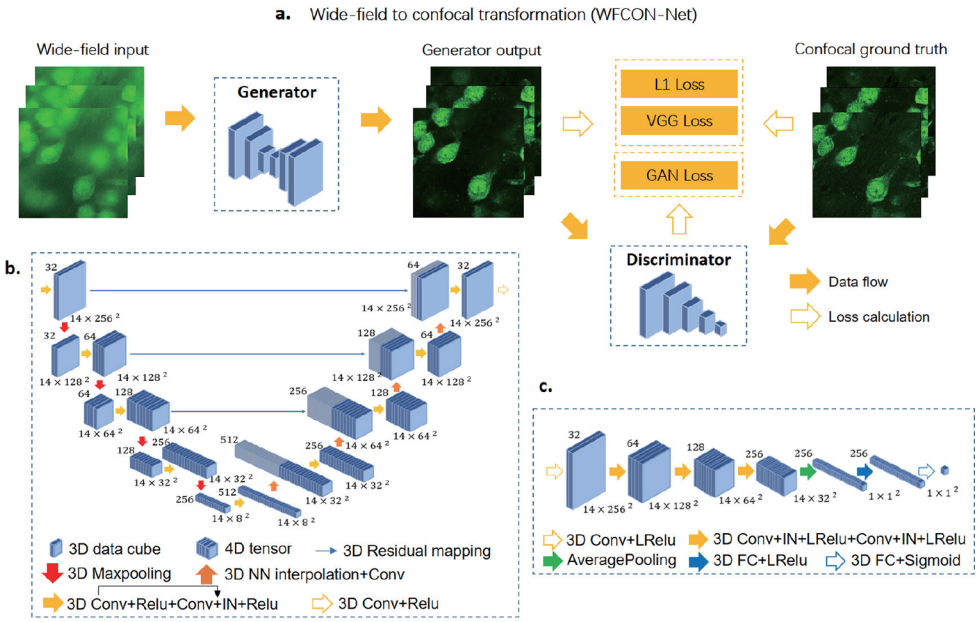

Fig. 2. Overview of the WFCON-Net architecture. **a)**. The WFCON-Net is a 3D GAN-based deep neural network, consisting of a generator and a discriminator. The generator takes 3D stacked wide-field fluorescent images as input, and outputs corresponding 3D confocal z-stack images in a single inference. The discriminator is learned to distinguish confocal ground truth from the prediction produced by the generator, with additional GAN losses. The use of the discriminator encourages the generator to predict confocal z-stacks with high accuracy, providing a good match to ground truth images. **b)**. Network architecture of the generator. **c)**. Network architecture of the discriminator.

The generator is a modified 3D convolutional U-net [40], consisting of an encoder path followed by a decoder path. The encoder path contains four down-sampling blocks: a max-pooling layer, two 3x3 3D convolutional layers, an instance normalization layer [41] and two Relu activation layers [42]. The use of the normalization layer makes the training of thick samples with diversely distributed signals and strong scattering backgrounds stable. In the decoder path, the

Biomedical Optics EXPRESS

max-pooling layer is symmetrically replaced by a nearest neighbor interpolation layer followed by a convolution layer with stride 1. The nearest neighbor interpolation layer, in our case, encourages the upsampling with fewer checkerboard artifacts than transpose convolutions. Moreover, residual mappings are performed between each convolutional block to guarantee the gradient flow. The discriminator is the 3D version of that in [30]. All the convolution and normalization operations are implemented in a 3D manner to explore the inter-layer relation of the volumetric z-stack data.

Training and testing data acquisition. To train/test our network, we captured 39 pairs of wide-field and confocal z-stacks images (with the lateral size of 2048×2048), using a developed setting of Andor Dragonfly spinning-disk confocal microscope that contains both wide-field fluorescent image capture mode and confocal fluorescent image capture mode. These image pairs of different regions of interest (ROIs) were randomly selected with different structures and neuron densities, covering the characteristics of different parts of the tissue slice, and are split into 'training' (31 pairs) and 'testing' (8 pairs) sets. As a similar remark to [40], more data cannot significantly enhance reconstruction quality but at a price of computational burden. The z-stacks are scanned with a step size of 0.5μm. The number of scans varies from 35 to 76, depending on the distribution of fluorescence signals along the z-axis. The same objective (60×/1.4NA oil, Nikon) was used for both wide-field and confocal imaging, and the resulting pixel size (in the image plane) is 108.3nm.

During training, the z-stacks were randomly cropped into 256×256×12 3D data patches. The data patches were then augmented with random flips and rotations, and normalized to $[0, 1]$ before feeding to the network. The batch size was set to 24. We trained our network for 6000 iterations (equivalent ~60 epochs) using NVIDIA TitanXp GPU and it takes 3 days for training.

Accurate image registration. To ensure the reconstruction quality of thick samples under high magnification, accurate image registration (i.e. data preparation to get the per-pixel matched wide-field and confocal image pairs) is indispensable. However, severe background and decreased contrast in wide-field images make commonly used cross-correlation registration (or calculate SSIM value) prone to error. Therefore, we proposed a new registration method tailored for 3D that calculates the PSF between confocal and wide-field image stacks, which can learn the physical connections between two stacks and more robust to noise. Then we use this PSF to determine the lateral and axial shifts. We termed three-dimensional confocal images as x, wide-field images as y. For simplicity, we treat the confocal images as ground-truth of sample distribution, then $y = f * x + n$, where f is the PSF of wide-field images and noise n results from non-uniform system/model errors and randomness of measurement. To robustly recover PSF, we add a standard TV (total variation) constraint on the gradients of the recovered PSF. Therefore, we can formulate the objective function as:

$$f_{opt} = \arg\min_{f} \|fx - y\|_2^2 + \lambda \|\nabla f\|_1 + \mu \|1^T f - 1\|_2^2 \tag{2}$$

where the first term is the least-squares data fitting term, the second term is TV gradients penalty, the last term enforces the energy conservation constraint, i.e., $\Sigma_{m,n} f(m, n) = 1$. In this experiment, λ is set to 1 and μ is set to 10. We use optimal first-order primal-dual framework [43,44] to optimize this objective function to get optimal f:

$$\begin{aligned} g_{n+1} &= Prox_{\sigma F^*}(g_n + \sigma K \bar{f_n}) \\ f_{n+1} &= Prox_{\tau G}(f_n + \tau K^* g_{n+1}) \\ \bar{f_n} &= f_{n+1} + \theta(f_{n+1} - f_n) \end{aligned} \tag{3}$$

Where σ, θ and τ are hyper-parameters, K is gradient operator, $*$ denotes the convex conjugate, $Prox_{F^*}$ and $Prox_G$ are proximal operators for function F^* and G, exact formula of these two operators can be found in [43]. Once f has been determined, the relative shift between wide-field and confocal image can be accurately calculated by the shift of maximal intensity point of PSF f.

Biomedical Optics EXPRESS

There is no need to estimate PSF at test time for test samples, it is only needed in the training step. Accurate image registration can substantially ensure the reconstruction quality.

Sample preparation of immune-fluorescent stained mouse brain slices. We demonstrated the performance of WFCON-Net on 40 microns thickness C57/B6 mouse brain slices obtained from Prof. Yichang Jia's lab, Tsinghua University. The neuron body and microglia of the brain slices are immune-fluorescent stained and the procedures are described as followed.

First, the brain slices were freshly obtained from Leica vibrating microtome 7000 (after perfusion-fixed with 4% paraformaldehyde in 1× PBS), and then incubated with permeate-blocking buffer (0.3% Triton-X100, 3%BSA in 1× PBS) at room temperature for 2 hours. After that, the slices were gently washed 5 times (5 min per wash) with washing buffer (0.05%Tween-20, 3%BSA in 1× PBS), incubated with NeuN antibody (Cell Signaling #94403, 100× dilution in washing buffer) and Iba1 antibody (Cell Signaling #17198, 100× dilution in washing buffer) at 4°C for 24 hours, protected from light. On the next day, the slices were gently washed 5 times again with washing buffer, then incubated with Alexafluor-488 labeled goat-anti-mouse antibody (Cell Signaling #4408, 200× dilution in washing buffer) and Alexafluor-555 labeled goat-anti-rabbit antibody (Cell Signaling #4413, 200× dilution in washing buffer) at 4°C overnight in the dark. On the third day, the slices were gently washed 5 more times with washing buffer, transferred on Superfrost Plus slides, mounted with 22 mm No.1.5 square coverslips, and ProLong Gold antifade mountant containing 2μg/mL DAPI. These prepared slices were protected from light and stored at 4°C before imaging. All the reagents, coverslips, and tissue slides were purchased from ThermoFisher if mentioned otherwise.

3. Results

3D confocal imaging of mouse brain slices using WFCON-Net. We first demonstrate our method on a mouse brain slice, as shown in Fig. 3(a). The wide-field inputs, as well as the prediction results, the corresponding confocal ground truths, and the difference maps of the selected region of interest (ROI) with different neuron densities/structures are shown in Fig. 3(b). We make use of the root mean square error (RMSE, the lower the better) and the structural similarity index measure (SSIM, the higher the better) to quantitatively evaluate the prediction accuracy. The average RMSE and SSIM of the testing datasets are 0.0575 and 0.7673. As we can see from the figure, the mice brain sample we have applied is a thick scattering sample (~40 microns), and the details of the wide-field images are completely overwhelmed by the scattering background. The predictions of such highly-scattering samples are significantly challenging than thinner samples (~ several microns) [30,33]. Our proposed GAN-based WFCON-Net, with 3D convolutional operations, can successfully reconstruct the high-contrast, high-resolution z-stack images from the wide-field captures, matching the confocal images well at the corresponding planes. The magnified y-z, x-z cross-sections of the image stacks (span 12μm in the z-direction), in Fig. 3(c) and (d), demonstrate the reconstruction accuracy across z-axis. The performance for the areas with denser neuron accumulation (such as ROI2) degrades slightly because of the more rigorous scattering background, leading to a larger reconstruction RMSE and a lower SSIM.

Our network also shows good generalization ability to unseen data. Without retraining, we tested the model with images captured from another two neuron slices, which have obviously different levels of background and scattering. Images of one slice have less background (Fig. 4(a,b)) while another suffer from more severe background (Fig. 4(c,d)). The images are shown in grayscale to clarify the level of background noise. We also tested cell/structure type (DAPI) different from training data (GFP) in Fig. 4(b,d)). The reconstructed images show good background suppression capability with little artifacts. Images of GFP channel (Fig. 4(a,c)), neuron cell body) exhibit higher accuracy than DAPI channel (Fig. 4(b,d)), nucleus) as the same channel used for the training. As images with more severe backgrounds share more similarities with the training set, their inference results are better than those images with less background.

Biomedical Optics EXPRESS

The average RMSE and SSIM of the testing datasets are 0.0695/0.7153 and 0.0544/0.7501 for GFP channel and 0.1106/0.6712 and 0.0923/0.6527 for DAPI channel. Samples with totally different structures will exhibit more artifacts, as the diversity of training data is limited. Further enhanced generalization ability can be envisioned as more data used for training.

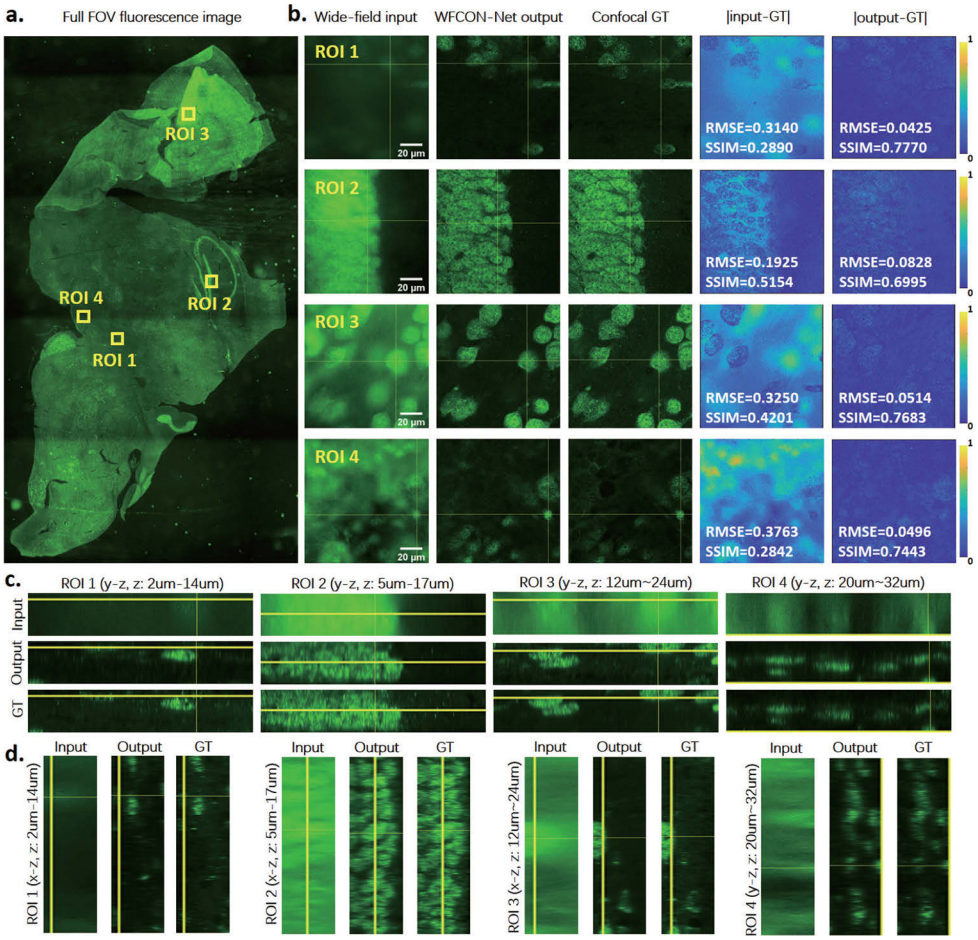

Fig. 3. 3D confocal imaging of mouse brain slice using WFCON-Net. a). The full field of view (FOV) lateral fluorescent image of a mouse brain slice. The full image is stitched together from 15 sub-images (for different parts of the slice) captured using a 4×/0.2 objective lens. **b**). Digital confocal predictions using WFCON-Net for 4 different ROIs with different neuron densities/structures. The wide-field inputs and confocal ground truths are also shown for comparison. The predicted confocal images match the ground truths (scanned using a confocal microscope) very well at corresponding x-y planes. The root mean square error (RMSE, the lower the better) and the structural similarity index measure (SSIM, the higher the better) are used here to quantitatively evaluate the prediction accuracy. The error maps |output-GT| show that the reconstruction qualities are slightly degraded with the increased neuron aggregation. **c**). Magnified views of y-z cross sections in 4 ROIs same to **b** show accurate three-dimensional confocal image reconstruction. The cross-sectional images span 12μm in the z-direction (with a step size of 0.5μm). The z-axis spans of those images are explicitly labeled. **d**). Magnified views of x-z cross sections in 4 ROIs same to **b**. The full-stack results span over 38μm (see **Visualization 1** and **Visualization 2**).

Biomedical Optics EXPRESS

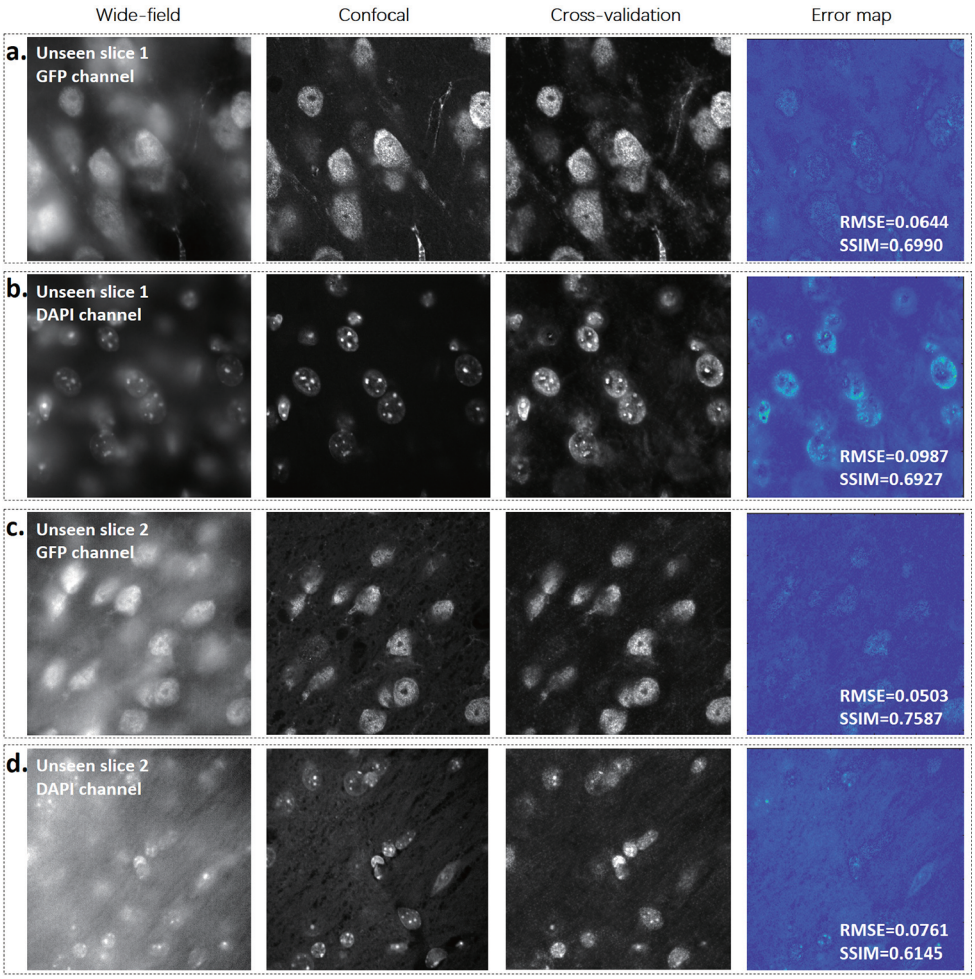

Fig. 4. Ablation study on model generalization ability. To show the generalization ability of our trained model, we tested it on two different neuron slices with different levels of background noise. Results for the green channel (GFP) and the blue channel (DAPI) are shown. **a,b)**. Results for the neuron slice with fewer scattering background than the training data. **c,d)**. Results for the neuron slice with more severe scattering background than the training data.

The image quality of inputs together with outputs are affected by the thickness of the sample. As axial depth increases, the sharpness of confocal ground-truth images and the reconstructions are both deteriorated. However, as shown in Fig. 5(a), the RMSE to the confocal ground-truth remains stable with different depths, which confirms the learning robustness of the proposed network. Furthermore, we measure the sharpness by calculating the gradient of image patches of size 32×32 by $\Delta I = |\Delta I_x| + |\Delta I_y|$. In each depth, we take the patch with the largest gradient as a sharpness measurement of signal, and the smallest one as a measurement of background. As shown in Fig. 5(b), the sharpness of the background keeps almost constant while the sharpness of confocal images is degraded along with depth. Although in superficial layers the confocal images outperform the network outputs, they continuously lose superiority when image deeper because the network can take advantage of a priori knowledge learned from shallower layer, which is a great advantage of our deep learning enabled digital confocal microscope.

Biomedical Optics EXPRESS

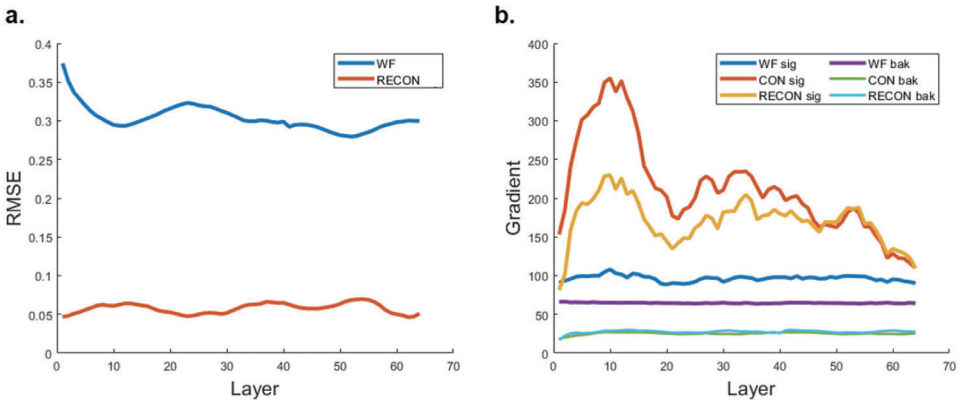

Fig. 5. Quantitative analysis of the reconstruction accuracy at different depths. a). The average RMSE of the wide-field images (WF) and the reconstructed confocal images (RECON) changes with depth. **b).** The average sharpness of the signal regions and the background regions in the wide-field images (WF), the confocal ground-truth images (CON) and the reconstructed confocal images (RECON), respectively.

WFCON-Net outperforms 2D convolutional networks for thick-sample prediction. Next, we compare our method with other 2D convolutional GAN-based networks. The prediction results for DeepZ+ [30], the 2D version of WFCON-Net and our 3D WFCON-Net are shown in Fig. 6. We implemented the DeepZ+ propagation algorithm as proposed in [30], which takes the single-layer wide-field image as input. The algorithm propagates the single wide-field image to predict multi-layer confocal z-stack images. Such propagation works well on the thin BPAEC microtubule structures, but fails to generate correct images for thick mice brain samples (Fig. 6(b)). We also investigated the 2D WFCON-Net by replacing the 3D convolutional blocks with 2D convolutions. Figure 6(c) and (d) display the predicted confocal images of 2D WFCON-Net and 3D WFCON-Net. The yellow boxes show a magnified view of the cell body and the insets mark the intensity profile in those specific regions. We plot the isosurface of the 3D spectrum of these image stacks in the last row to quantitatively measure the enhancement of the Fourier coverage after the network. The enhancement is more obvious in lateral than in axial, because of the limited sampling rate in the axial dimension. The confocal images in our case have asymmetric spectrum in lateral, but they are corrected by the network due to data augmentation such as rotation and flip. The spectrum also reveals the high-frequency artifacts of 2D network that otherwise can only be observed with a very magnified view. The spectrum of DeepZ+ results has the smallest Fourier coverage because it has only one input 2D image as the source of information. To summarize, our 3D WFCON-Net benefits from the stronger representation ability and inter-layer correlation information, outperforming the 2D methods in confocal predictions with less blur, richer details and higher resolution.

Training with GAN and VGG loss. Moreover, we introduce the perceptual VGG loss together with GAN loss to train our network. VGG loss is prevalent in natural image super-resolution as it can enhance the high frequency details that are filtered in low-resolution images, making the image more realistic [45]. On the contrary, conventional PSNR oriented loss tends to smooth the reconstruction result [39]. By adding the VGG loss, the reconstruction details are well preserved (especially in the overexposed area) and the background noises are well suppressed (Fig. 7). Similarly, in the lateral frequency plane, the enhancement of Fourier coverage is obvious. After taking normalization to compensate the intensity mismatch, we quantitatively calculate the sum of the lateral spectrum within the red box, which is in consistency with our conclusion that the use of the perceptual loss encourages high-quality, high-resolution predictions, matching the

Biomedical Optics EXPRESS

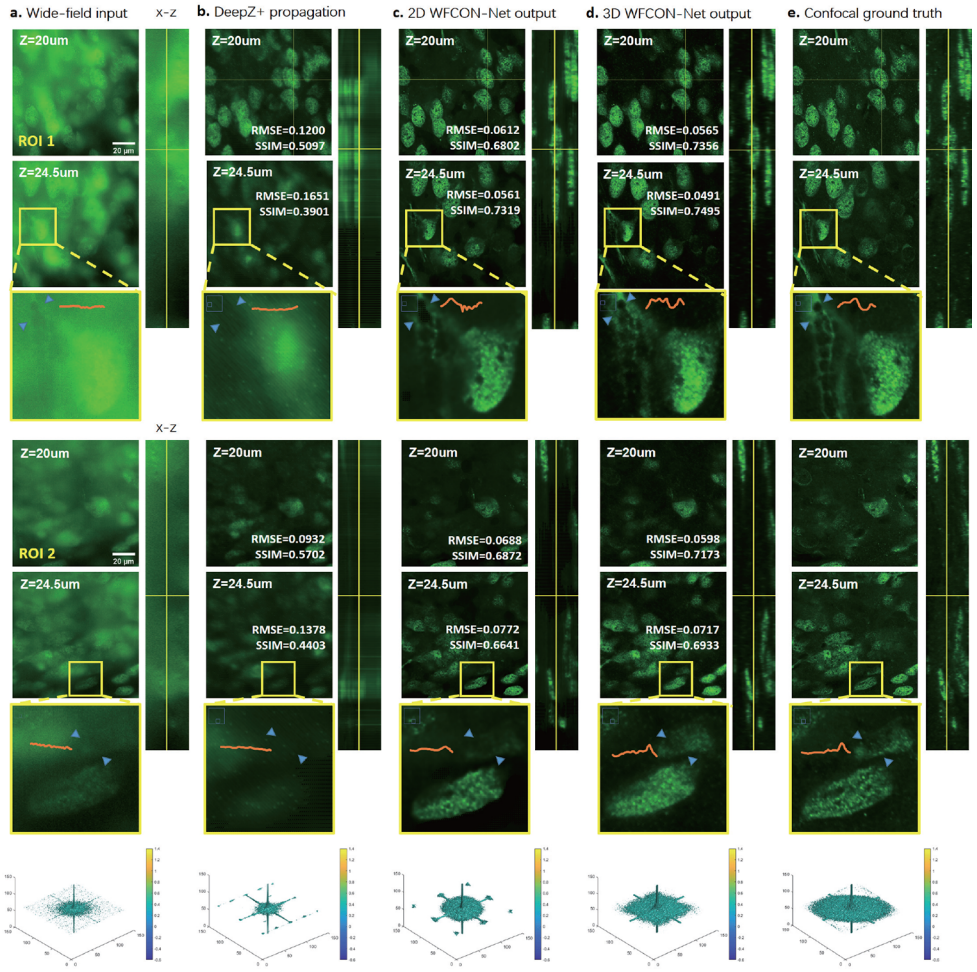

Fig. 6. Comparison with other wide-field to confocal cross-modality methods. a). Wide-field input images of two ROIs. **b**). Propagated 3D confocal images using DeepZ+ [30]. The DeepZ+ takes the single-layer wide-field image (eg. layer z=20μm) as input, and outputs the propagated 3D confocal z-stack images (only layer z=24.5μm is shown). The propagation fails to generate accurate predictions for thick samples (due to the distributed fluorescence signals in multiple layers and the strong scattering background). **c**). Predicted confocal images using 2D WFCON-Net. The 2D WFCON-Net (with 2D Conv blocks) takes the wide-field z-stack images as input, and predicts the corresponding confocal images in a layer-to-layer manner. **d**). Predicted confocal images using WFCON-Net. The WFCON-Net (with 3D Conv blocks) benefits from the stronger representation ability and inter-layer correlation information, and thus surpasses 2D methods in confocal predictions with higher accuracy and richer details, which can be verified by the line profile marked by two triangular arrows in the insets of images. **e).** Confocal ground truth images. For all the images, the x-y images at z=20μm, 24.5μm and their corresponding x-z cross-sections are shown. The last row displays the isosurface of the 3D spectrum of image stacks.

corresponding confocal images well. The result also demonstrates that different image generation tasks share some feature similarities.

PSF-based registrations. Image registration is one of the main issues for learning-based cross-modality image reconstruction. The greater the degree of scattering and noise, the higher

Biomedical Optics EXPRESS

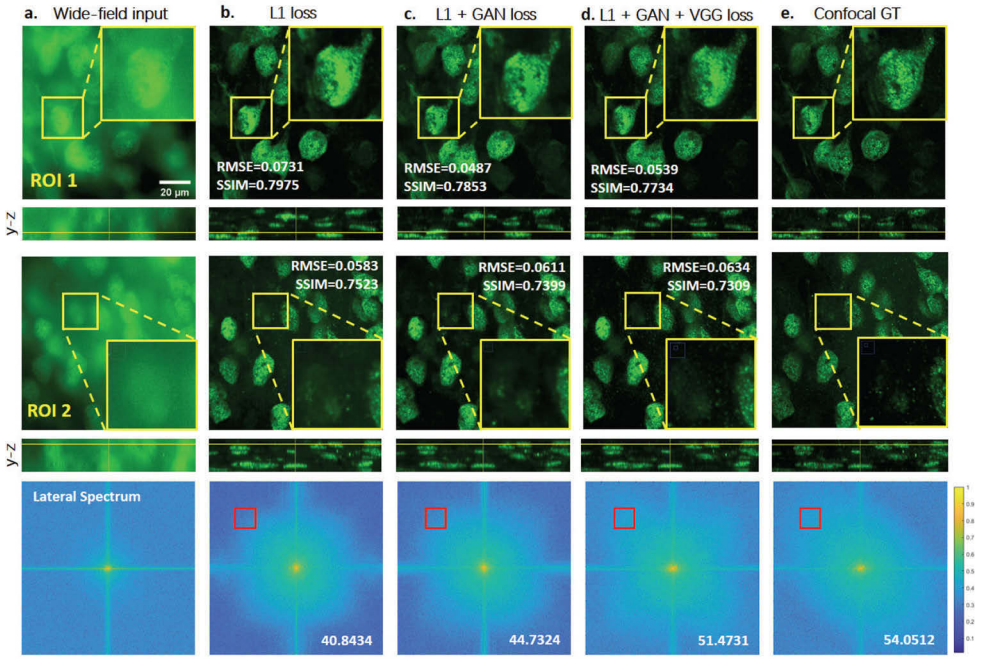

Fig. 7. Training with GAN and VGG loss. We used the sum of L1 loss, VGG loss and GAN loss as our loss function during training. The use of perceptual loss (VGG loss) encourages predictions with higher quality and richer details. **a)**. Wide-field input stack of two different ROIs, the last row shows the lateral frequency plane of the image stacks. **b)**. Predicted confocal images of the network trained with only L1 loss [33]. **c)**. Predicted confocal images of the network trained with L1 loss and GAN loss [30]. **d)**. Our method, trained with a combination of L1 loss, VGG loss and GAN loss. **e)**. Confocal ground truth images. Although the use of non-PSNR oriented perceptual loss will slightly affect SSIMs and RMSEs, but as shown in the magnified views, our method can reconstruct more details (especially in the overexposed areas – ROI1) and suppress more background noises (ROI2).

the difficulty for registration. As shown in Fig. 8(a), fluorescent structures in thick sample are more ambiguous than thin sample, resulting in the decrease of the accuracy of registration and the degradation of the reconstruction performance. We can observe that the cross-correlation curve along the z-axis is flatter than the PSF curve (Fig. 8(b)), especially in thick samples. This flatness causes difficulty in identifying the actual shift between image pairs. By calculating the PSF between 3D wide-field and confocal image pairs, we find the physical connections between two modalities and the registration process becomes more accurate, reconstructed details are thus well preserved (Fig. 8(c)). The registration error of the PSF estimation method occurs only when the shift between wide-field and confocal images falls in the middle of two integer numbers, corresponding to two peaks with approximate height as shown in the left plot of Fig. 8(b). Hence, the maximum theoretical error is nearly half of the minimum z-step size. The PSF registration error can therefore be further reduced with a smaller z-step size of training data.

Ablation study on the number of wide-field input layers. Our method requires capturing the whole stack of wide-field images by scanning along the z-direction. To accommodate the applications that demand faster data acquisition, we investigated the performance of our network trained with the whole stacks but tested with fewer images as input. Specifically, we subsample the layers of captured z-stacks at an interval of 2 layers (downsample 2×) and 4 layers (downsample 4×), and interpolate the missing layers to the original size before feeding

Biomedical Optics EXPRESS

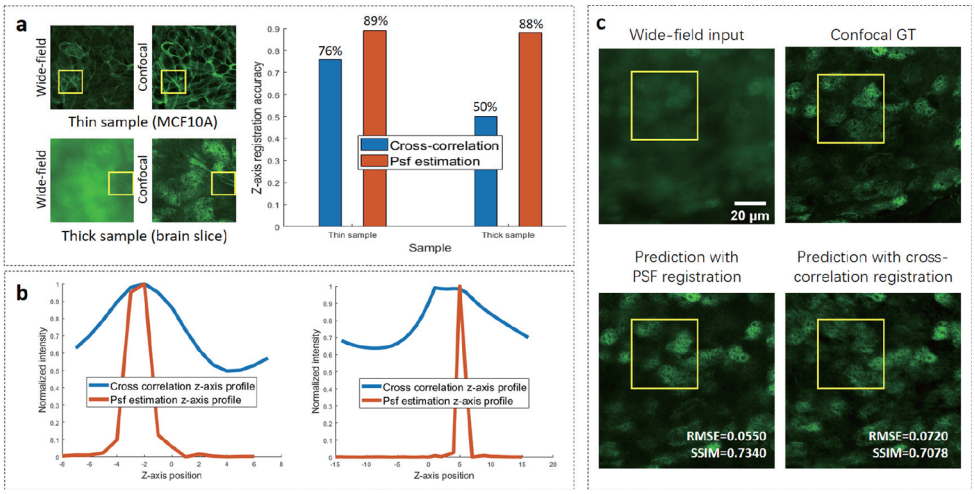

Fig. 8. Image registration with estimated PSF. We registered images by estimating the 3D point spread functions (PSF) from paired wide-field and confocal z-stack images. The peak of the calculated PSF profile is then used to determine the lateral and axial shifts between unregistered images. **a)**. We compare the PSF registration method with the cross-correlation method for a thin sample (MCF10A, 3 microns thickness) and a thick sample (brain slice, 38 microns thickness), respectively. The PSF registration (orange) outperforms the cross-correlation method (blue) in higher z-axis accuracy, especially for the thick samples. **b)**. Profiles along the z-axis of cross-correlation and calculated PSF for the thin sample (left) and the thick sample (right). The sharp peak of PSF benefits the accurate registration and is robust to noises across wide-field and confocal images. **c)**. WFCON-Net predictions with and without PSF registered training data. The PSF registration method improves the reconstruction quality. The ground-truth registrations are manually aligned.

them into the network. The downsampled z-stacks come with an equivalent z-step of 1μm and 2μm, respectively (original z-step ~0.5μm). Figure 9 shows the reconstruction results with 2× downsampling, 4× downsampling, and without downsampling (without retraining the network). We can see that the model trained with the original data (no downsampling) works well for the 2× downsampling, and is slightly degraded for 4× downsampled z-stacks. The spacing of 4× downsampled images in z-axis is same to the spacing Huang et al. have demonstrated using 8 micronsC. *elegans* data [35], but we show good results on samples with much more complicated structures and unwanted backgrounds. The degradation is mainly caused by the gap between captured input layers and the interpolated layers. Higher reconstruction quality can be obtained if we retrain the network with the interpolation process included in the pipeline. The average RMSE and SSIM of the testing dataset for without/2×/4× downsampling are 0.0575/0.0598/0.0657 and 0.7673/0.7343/0.6652, respectively.

Biomedical Optics EXPRESS

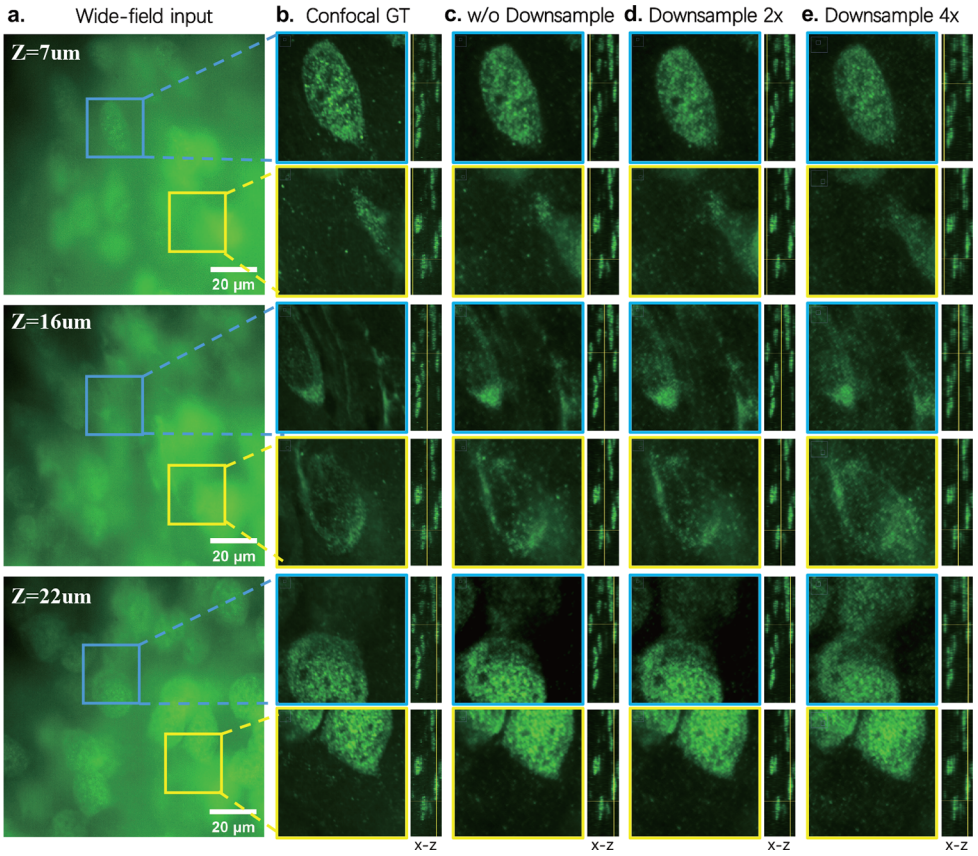

Fig. 9. Ablation study on the number of wide-field input layers (along z). We tested our algorithm with the different levels of input layers by downsampling the wide-field z-stacks, without retraining the network. The whole z-stacks are interpolated from the downsampled data before feeding to our network. **a)**. Wide-field inputs at different depths. Two specific regions of interest (bounded by the blue and yellow boxes) are enlarged to show details. The x-z cross-sections are also shown for localization. **b)**. Confocal (ground truth, GT) images of the two enlarged regions. **c)**. WFCON-Net reconstructed images without downsampling. **d)**. WFCON-Net reconstructed images with 2× downsampled input and interpolation. The results are comparable to the results without downsampling. **e)**. WFCON-Net reconstructed images with 4× downsampled input and interpolation. The reconstruction accuracy degrades slightly and mainly in out-of-focus layers, since these layers are more vulnerable to the background signals originating from the in-focus layers.

4. Discussion

We provide a 3D GAN-based neural network to generate the confocal images from the wide-field images. To the best of our knowledge, we are the first to deal with wide-field images of a thick and complex sample, in which input image quality is severely degraded by scattering and background noise. By using GAN-based 3D U-Net with additional residual mapping, normalization layer and VGG loss, together with accurate image registration, the reconstructed details are repaired from wide-field images. Our model has shown good generalization ability, either channel generalization or sample generalization, and can also be used with downsampled inputs in z-axis. Our experimental results demonstrate generalization ability to handle unseen data, stability in the reconstruction results, high spatial resolution even when imaging thick (~40

Biomedical Optics EXPRESS

microns) highly-scattering samples. We believe that such learning-based-microscopes have the potential to democratize scientific imaging bringing confocal quality imaging to every lab that has a wide-field microscope.

Our method is currently limited by the requirement of the training dataset to achieve the best performance. In some scenarios, like in-vivo neuron activity imaging of moving mice, the ground truth confocal or two-photon images are hard to acquire and register to the wide-field inputs. If we use the model trained on another sample, the inference result might be degraded. To deal with this problem, we can generate images from simulations [46] or using unpaired GAN framework as CycleGAN [47]. Measuring the network reliability [17] or discovering a more powerful registration method that can correct rotation and magnification errors also need further investigations.

Funding. National Natural Science Foundation of China (61620106005, 61827804); National Science Foundation (CAREER Award 1652633, SaTC Award 1801372).

Acknowledgement. We would like to thank Yichang Jia and Xu Zhang from Tsinghua University for their help on preparing the samples. We would also like to thank Xu Chen from Tsinghua Shenzhen International Graduate School for preparing the visualization video. In addition, we thank Yanli Zhang and Yalan Chen from Technology Center for Protein Sciences of Tsinghua University for assistance of using microscopes.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request. The code will be released in the future.

References

1. T. Wilson, "Confocal microscopy," San Diego (1990).
2. J. Pawley, *Handbook of Biological Confocal Microscopy*, vol. 236 (Springer Science & Business Media, 2006).
3. F. Helmchen and W. Denk, "Deep tissue two-photon microscopy," Nat. Methods **2**(12), 932–940 (2005).
4. P. T. So, C. Y. Dong, B. R. Masters, and K. M. Berland, "Two-photon excitation fluorescence microscopy," Annu. Rev. Biomed. Eng. **2**(1), 399–429 (2000).
5. J.-H. Park, W. Sun, and M. Cui, "High-resolution in vivo imaging of mouse brain through the intact skull," Proc. Natl. Acad. Sci. **112**(30), 9236–9241 (2015).
6. R. Lu, W. Sun, Y. Liang, A. Kerlin, J. Bierfeld, J. D. Seelig, D. E. Wilson, B. Scholl, B. Mohar, M. Tanimoto, M. Koyama, D. Fitzpatrick, M. B. Orger, and N. Ji, "Video-rate volumetric functional imaging of the brain at synaptic resolution," Nat. Neurosci. **20**(4), 620–628 (2017).
7. I. N. Papadopoulos, J.-S. Jouhanneau, N. Takahashi, D. Kaplan, M. Larkum, J. Poulet, and B. Judkewitz, "Dynamic conjugate f-sharp microscopy," Light: Sci. Appl. **9**(1), 110 (2020).
8. W. Zong, R. Wu, S. Chen, J. Wu, H. Wang, Z. Zhao, G. Chen, R. Tu, D. Wu, Y. Hu, Y. Xu, Y. Wang, Z. Duan, H. Wu, Y. Zhang, J. Zhang, A. Wang, L. Chen, and H. Cheng, "Miniature two-photon microscopy for enlarged field-of-view, multi-plane and long-term brain imaging," Nat. Methods **18**(1), 46–49 (2021).
9. B.-C. Chen, W. R. Legant, K. Wang, L. Shao, D. E. Milkie, M. W. Davidson, C. Janetopoulos, X. S. Wu, I. Hammer, A. John, Z. Liu, B. P. English, Y. Mimori-Kiyosue, D. P. Romero, A. T. Ritter, J. Lippincott-Schwartz, L. Fritz-Laylin, R. D. Mullins, D. M. Mitchell, J. N. Bembenek, A.-C. Reymann, R. Boehme, S. W. Grill, J. T. Wang, G. Seydoux, U. S. Tulu, D. P. Kiehart, and E. Betzig, "Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution," Science **346**(6208), 1257998 (2014).
10. P. J. Keller, A. D. Schmidt, J. Wittbrodt, and E. H. Stelzer, "Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy," Science **322**(5904), 1065–1069 (2008).
11. R. M. Power and J. Huisken, "A guide to light-sheet fluorescence microscopy for multiscale imaging," Nat. Methods **14**(4), 360–373 (2017).
12. L. Gao, L. Shao, C. D. Higgins, J. S. Poulton, M. Peifer, M. W. Davidson, X. Wu, B. Goldstein, and E. Betzig, "Noninvasive imaging beyond the diffraction limit of 3d dynamics in thickly fluorescent specimens," Cell **151**(6), 1370–1385 (2012).
13. K. McDole, L. Guignard, F. Amat, A. Berger, G. Malandain, L. A. Royer, S. C. Turaga, K. Branson, and P. J. Keller, "In toto imaging and reconstruction of post-implantation mouse development at the single-cell level," Cell **175**(3), 859–876.e33 (2018).
14. Y. Yue, W. Zong, X. Li, J. Li, Y. Zhang, R. Wu, Y. Liu, J. Cui, Q. Wang, Y. Bian, X. Yu, Y. Liu, G. Tan, Y. Zhang, G. Zhao, B. Zhou, L. Chen, W. Xiao, H. Cheng, and A. He, "Long-term, in toto live imaging of cardiomyocyte behaviour during mouse ventricle chamber formation at single-cell resolution," Tech. rep., Nature Publishing Group (2020).
15. Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," Optica **4**(11), 1437–1443 (2017).
16. H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydın, L. A. Bentolila, C. Kural, and A. Ozcan, "Deep learning enables cross-modality super-resolution in fluorescence microscopy," Nat. Methods **16**(1), 103–110 (2019).

Biomedical Optics EXPRESS

17. Y. Xue, S. Cheng, Y. Li, and L. Tian, "Reliable deep-learning-based phase imaging with uncertainty quantification," Optica **6**(5), 618–629 (2019).
18. H. Zhang, Y. Zhao, C. Fang, G. Li, M. Zhang, Y.-H. Zhang, and P. Fei, "Exceeding the limits of 3d fluorescence microscopy using a dual-stage-processing network," Optica **7**(11), 1627–1640 (2020).
19. C. Bai, X. Yu, T. Peng, C. Liu, J. Min, D. Dan, and B. Yao, "3d imaging restoration of spinning-disk confocal microscopy via deep learning," IEEE Photonics Technol. Lett. **32**(18), 1131–1134 (2020).
20. A. Goy, K. Arthur, S. Li, and G. Barbastathis, "Low photon count phase retrieval using deep learning," Phys. Rev. Lett. **121**(24), 243902 (2018).
21. M. Weigert, U. Schmidt, T. Boothe, A. Mueller, A. Dibrov, A. Jain, B. Wilhelm, D. Schmidt, C. Broaddus, S. Culley, M. Rocha-Martins, F. Segovia-Miranda, C. Norden, R. Henriques, M. Zerial, M. Solimena, J. Rink, P. Tomancak, L. Royer, F. Jug, and E. W. Myers, "Content-aware image restoration: pushing the limits of fluorescence microscopy," Nat. Methods **15**(12), 1090–1097 (2018).
22. W. Ouyang, A. Aristov, M. Lelek, X. Hao, and C. Zimmer, "Deep learning massively accelerates super-resolution localization microscopy," Nat. Biotechnol. **36**(5), 460–468 (2018).
23. E. Nehme, D. Freedman, R. Gordon, B. Ferdman, L. E. Weiss, O. Alalouf, T. Naor, R. Orange, T. Michaeli, and Y. Shechtman, "Deepstorm3d: dense 3d localization microscopy and psf design by deep learning," Nat. Methods **17**(7), 734–740 (2020).
24. X. Li, J. Dong, B. Li, Y. Zhang, Y. Zhang, A. Veeraraghavan, and X. Ji, "Fast confocal microscopy imaging based on deep learning," in *2020 IEEE International Conference on Computational Photography (ICCP)*, (IEEE, 2020), pp. 1–12.
25. Y. Wu, Y. Rivenson, Y. Zhang, Z. Wei, H. Günaydin, X. Lin, and A. Ozcan, "Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery," Optica **5**(6), 704–710 (2018).
26. H. Pinkard, Z. Phillips, A. Babakhani, D. A. Fletcher, and L. Waller, "Deep learning for single-shot autofocus microscopy," Optica **6**(6), 794–797 (2019).
27. C. Ounkomol, S. Seshamani, M. M. Maleckar, F. Collman, and G. R. Johnson, "Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy," Nat. Methods **15**(11), 917–920 (2018).
28. E. M. Christiansen, S. J. Yang, D. M. Ando, A. Javaherian, G. Skibinski, S. Lipnick, E. Mount, A. O'Neil, K. Shah, A. K. Lee, P. Goyal, W. Fedus, R. Poplin, A. Esteva, M. Berndl, L. L. Rubin, P. Nelson, and S. Finkbeiner, "In silico labeling: predicting fluorescent labels in unlabeled images," Cell **173**(3), 792–803.e19 (2018).
29. Y. Rivenson, H. Wang, Z. Wei, K. de Haan, Y. Zhang, Y. Wu, H. Gunaydin, J. E. Zuckerman, T. Chong, A. E. Sisk, L. M. Westbrook, W. D. Wallace, and A. Ozcan, "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning," Nat. Biomed. Eng. **3**(6), 466–477 (2019).
30. Y. Wu, Y. Rivenson, H. Wang, Y. Luo, E. Ben-David, L. A. Bentolila, C. Pritz, and A. Ozcan, "Three-dimensional virtual refocusing of fluorescence microscopy images using deep learning," Nat. Methods **16**(12), 1323–1331 (2019).
31. X. Zhang, Y. Chen, K. Ning, C. Zhou, Y. Han, H. Gong, and J. Yuan, "Deep learning optical-sectioning method," Optics Express (2018).
32. S. Lim, H. Park, S. Lee, S. Chang, B. Sim, and J. C. Ye, "Cyclegan with a blur kernel for deconvolution microscopy: Optimal transport geometry," IEEE Trans. Comput. Imaging **6**, 1127–1138 (2020).
33. K. Ning, X. Zhang, X. Gao, T. Jiang, H. Wang, S. Chen, A. Li, and J. Yuan, "Deep-learning-based whole-brain imaging at single-neuron resolution," Biomed. Opt. Express **11**(7), 3567–3584 (2020).
34. J. Lim, A. B. Ayoub, and D. Psaltis, "Three-dimensional tomography of red blood cells using deep learning," Adv. Photonics **2**(02), 1 (2020).
35. L. Huang, H. Chen, Y. Luo, Y. Rivenson, and A. Ozcan, "Recurrent neural network-based volumetric fluorescence microscopy," Light: Sci. Appl. **10**(1), 62 (2021).
36. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems **27**, (2014).
37. M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784 (2014).
38. X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, (2017), pp. 2794–2802.
39. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, (Springer, 2016), pp. 694–711.
40. Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, (Springer, 2016), pp. 424–432.
41. D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," arXiv preprint arXiv:1607.08022 (2016).
42. X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, (2011), pp. 315–323.
43. F. Heide, M. Rouf, M. B. Hullin, B. Labitzke, W. Heidrich, and A. Kolb, "High-quality computational imaging through simple lenses," ACM Trans. Graph. **32**(5), 1–14 (2013).
44. A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," J. Math. Imaging Vis. **40**(1), 120–145 (2011).

Biomedical Optics EXPRESS

45. C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang, "Photo-realistic single image super-resolution using a generative adversarial network," IEEE Computer Society (2016).
46. Z. Pengcheng, S. L. Resendez, R. R. Jose, J. C. Jimenez, S. Q. Neufeld, G. Andrea, F. Johannes, E. A. Pnevmatikakis, G. D. Stuber, and H. a. Rene, "Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data," eLife **7**, e28728 (2018).
47. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, (2017), pp. 2223–2232.